

SPEAKER RECOGNITION METHOD BASED ON STRUCTURED SPEAKER

MODELING AND A "PICKMAX" SCORING TECHNIQUE

Field of the Invention

The present invention generally relates to score calculation and normalization in a
5 framework of speaker recognition with phonetically structured speaker models.

Background of the Invention

Typically, in speaker recognition systems, a sample of the voice properties of a
target speaker is taken and a corresponding voice print model is built. In order to improve
system robustness against impostors in a "verification" mode, it is also typical for a large
10 number of non-target speakers (*i.e.*, "background speakers") to be analyzed, pre-stored,
and then used to normalize the voice-print likelihood score of the target speakers.

The voice analysis can be conducted at various levels of phonetic detail, ranging
from global (phoneme-independent) models to fine phonemic or subphonemic levels.
With several such levels in a system, a problem arises as to how to combine scores from
15 different levels. Combining scores from different levels may be important since it may not
always be possible to obtain data at the phonemic level. Particularly, while it is recognized
that the voice patterns of a speaker vary with phonemes (or sounds), and are thus better

distinguished by models that are created for individual phonemes, it is sometimes the case that the training data will be sparse. In this case, not all of the phoneme models can be created in a robust way (*i.e.*, in terms of statistical robustness) and therefore have to be combined with models created on a higher level of coarseness (or granularity), such as on
5 broad classes of phonemes (vowels, plosives, fricatives etc.) or on phoneme-independent models, whose robustness is higher. Conventionally, this combination is achieved as a linear interpolation of the model scores from individual granularity levels in a method known as the "back-off" method. A discussion of the "back-off" method can be found in F. Jelinek, "Statistical Methods for Speech Recognition" (MIT Press 1998, ISBN
10 0262100665). However, this method, as well as other conventional methods, have often been found to be inadequate in providing effective speech verification capabilities.

Accordingly, a need has been recognized in connection with providing a system that adequately and effectively combines scores from the individual levels while avoiding other shortcomings and disadvantages associated with conventional arrangements.

15 **Summary of the Invention**

The present invention broadly contemplates, in accordance with at least one presently preferred embodiment, the calculation of scores in such a way that the total likelihood is a weighted sum of the likelihood of all phonetic units at all levels of phonetic

granularity (model grains), and that the weights are derived in such a way that the determination of the robustness and significance of the individual model grains is approached with emphasis.

A particular manner of designing these weights on-the-fly is contemplated herein
5 that takes the actual likelihoods of the test utterance into account and allows for determining the level of distinction as well as the phonetic correspondence on-the-fly using a maximum-likelihood criterion for the individual feature vectors. Apart from the improved accuracy, such an arrangement permits a significant reduction in computation during the verification stage since there is no need for explicit phonetic labeling of the test
10 utterance.

It should be understood that the present invention, in broadly contemplating speaker "recognition", encompasses both speaker verification and speaker identification. With regard to "identification", this may be understood as a task of recognizing a previously enrolled speaker based solely on a test utterance (*i.e.*, no additional identity
15 claims are provided, as opposed to verification). The identification result is the recognized speaker's identity (name, number, etc.; as opposed to the binary "accept/reject" result with verification). Typically, for identification, no background population is necessary for normalization. The task is posed as statistical classification problem and typically solved

using a maximum-likelihood classifier. Identification processes contemplated herein address the calculation of the basis likelihood of a frame given a model (just as in the verification mode). Practical applications for identification include automatic user recognition for adaptation. For instance, a speech-enabled application, *e.g.*, a PC-desktop
5 or a personal email assistant over the telephone, can recognize which user is requesting a service without explicitly requiring his/her name or ID.

In one aspect, the present invention provides a method of providing speaker recognition, the method comprising the steps of: providing a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of
10 phonetic detail; receiving an identity claim; ascertaining whether the identity claim corresponds to the target speaker model; the ascertaining step comprising the steps of determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolving the at least one likelihood value to obtain a likelihood score.

15 In another aspect, the present invention provides an apparatus for of providing speaker recognition, the apparatus comprising: a target speaker model generator for generating a model corresponding to a target speaker, the model being resolved into at least one frame and at least one level of phonetic detail; a receiving arrangement for

receiving an identity claim; a decision arrangement for ascertaining whether the identity claim corresponds to the target speaker model; the decision arrangement being adapted to determine, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolve the at least one likelihood value to obtain a
5 likelihood score.

Furthermore, the present invention provides in another aspect a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for providing speaker recognition, the method comprising the steps of: providing a model corresponding to a target speaker, the model
10 being resolved into at least one frame and at least one level of phonetic detail; receiving an identity claim; ascertaining whether the identity claim corresponds to the target speaker model; the ascertaining step comprising the steps of determining, for each frame and each level of phonetic detail of the target speaker model, a non-interpolated likelihood value; and resolving the at least one likelihood value to obtain a likelihood score.

15 For a better understanding of the present invention, together with other and further features and advantages thereof, reference is made to the following description, taken in conjunction with the accompanying drawings, and the scope of the invention will be pointed out in the appended claims.

Brief Description of the Drawings

Figure 1 illustrates an example of a structure speaker model (voice-print) with three levels and a variable number of units on each level.

Figure 2 illustrates a speaker verification system with the "Pickmax" scoring and
5 structure speaker models.

Figure 3 illustrates a speaker identification system using the "Pickmax" scores and a maximum-likelihood classifier.

Description of the Preferred Embodiments

The target as well as the background speaker population (used for cohort-based
10 score normalization) are enrolled into the system by creating their statistical models in the feature space. The enrollment utterances are preferably phonetically structured using a transcription engine or a phonetic labeler (for example, a ballistic decoder as described in copending and commonly assigned U.S. Patent Application Serial No. 09/015,150 or forced alignment as described copending and commonly assigned U.S. Patent Application
15 Serial No. 09/519,327).

Based on the labeling information, the data is preferably structured on predefined levels of phonetic detail into units, for instance, global level, phone-class level, and phone

level. It is to be noted, however, that the levels may not necessarily obey a top-down or bottom-up detail hierarchy as in the present example. Corresponding models are then preferably created for each of the units for a given speaker. These so-called structured models represent the speakers' voice-prints, as shown in Figure 1.

5 Thus, Figure 1 illustrates a structured speaker model 100 that may include statistical models of different "levels" as discussed above, for instance, a global level 102, a phone-class level 104 and a phone level 106. A global level 102 will preferably involve a model created from all feature vectors, a phone-class level 104 may preferably include models created for broad phonemic classes (*e.g.*, vowels, nasals, plosives, fricatives,
10 liquids etc.), while a phone level 106 may preferably include single phones (*e.g.*, "aa", "oh", "n", etc.).

The disclosure now turns to a process of verification in accordance with a presently preferred embodiment of the present invention, as described herebelow and as illustrated in Figure 2.

15 With regard to a conventional procedure against which at least one presently preferred embodiment of the present invention may be compared, let U denote a test utterance (203) that includes T feature vectors (frames) u_1, \dots, u_T , which utterance is to be verified based on a speaker's claimed identity 200c. In this connection, a "claim" refers to

an identification tag (such as an identification number, label, name, etc.) to which a speaker claims to correspond. A claimed identity, then, may be expressed the speaker (for example) as "my name is Jerry," or "my customer number is 1234". The existence of a claim is essential for the verification.

- 5 Given a structured model $M(i,j)$ for a speaker with $1 \leq i \leq L$ levels of detail and with $1 \leq j \leq K(i)$ units on the i -th level, the score (as log-probability) for the utterance is calculated in each level separately, whereby explicit labeling information is used to identify the corresponding phonetic unit that is to be used on each level:

Sub Pt
$$S(U|M) = \frac{1}{T} \sum_{i=1}^L \alpha_i \cdot \sum_{t=1}^T P(u_t | M(i, j, (i, t))) \quad (1)$$

- 10 where α_i is an interpolation constant for level i and $j(i,t)$ is the labeling information (unit) for level i and frame t . As examples of labelling information that could be used as $j(i,t)$, one might encounter, for instance, $j(1,1)=1$ and $j(2,1)=4$, meaning that in the time-frame $t=1$: on level=1 use unit number 1 (which might be for example the only model if the level is the "global" one), and on level=2 use unit number 4 (which might correspond to a class
- 15 of phonemes such as "fricatives.")

The formula (1) may now be generalized, in terms of weighing, by assigning specific weights to each of the units at each level (i.e. to each grain) as follows:

$$S(U | M) = \frac{1}{T} \sum_{i=1}^L \sum_{t=1}^T b_{i,j(i,t)} \cdot P(u_t | M\{i, j(i,t)\}) \quad (2)$$

with $b_{\{i,j(i,t)\}}$ denoting grain-specific weights that satisfy

$$\sum_{i=1}^L \sum_{j=1}^{K(i)} b_{ij} = 1 \quad (3)$$

The weights b may be derived in a way so as to emphasize a) grains whose training data amount was above average, thus, which are expected to be more robust, or b) grains which showed an above-average contribution to the performance measured on some development data set or c) grains that are significant with respect to the current test utterance, all subject to the constraint (3). The latter method is further refined below and an algorithm for determining the weights on-the-fly is described ("pickmax").

In a "pickmax" technique in accordance with an embodiment of the present invention (step 209), the likelihood score S for each of the structured models mentioned above is calculated as the average of the likelihoods of the T feature vectors which, in turn, are obtained as the maximum likelihoods computed over all units and all levels of the given speaker's structured model ("pickmax"). This corresponds to deriving the weights $b_{\{i,j\}}$ in (2) from the likelihood of the actual utterance frame at the time t based on all grains, as follows:

$$b_{ij} = 1 \text{ for } \{i, j\} = \arg \max_{1 \leq i \leq L, 1 \leq j \leq K(i)} P(u_t | M\{i, j\})$$

$$b_{i,j} = 0 \text{ otherwise}$$

Since there is only one such maximum (or only one is taken in cases of two or more numerically equal maxima) the constraint (3) is implicitly satisfied.

5 Thus, the formula (2) can be rewritten as:

$$S(U | M) = \frac{1}{T} \sum_{t=1}^T \max_{1 \leq i \leq L, 1 \leq j \leq K(i)} P(u_t | M(i, j)) \quad (4)$$

It is to be noted that the number of units on each level and the number of levels may vary across speakers, since there might be less data available from certain speakers, entailing the necessity of omitting certain units altogether. The scores calculated in (4)

10 will thus still be appropriate for such inter-speaker inconsistencies in the models. Unlike in equation (1), in equation (4) there is no labeling information and no need for interpolation constants which typically must be obtained from independent data sets and can be a source of “over-training.” By this, what is meant is that the additional constant must be determined on some data. If there is not enough data, this constant will be

15 determined in too specific a manner with respect to the training and will not be sufficiently

general. It is to be noted that the score calculation (2) and (4) is suitable for (and, without modifications can be applied to) both the speaker verification and identification problems, as well as other related tasks, such as speaker segmentation and tracking.

The final verification is preferably carried out by calculating the log-likelihood ratio
5 (also step 209), for instance, as

$$L = S(U | M) - \frac{1}{C} \sum_{i=1}^C S(U | BG_i) \quad (5)$$

where M denotes the target model and BG_i the i -th background (cohort) model. The value of L , or the "final score" or "discriminant", preferably undergoes a threshold decision (at 211) to either accept or reject (213) the hypothesis that the utterance was spoken by the
10 target speaker. Alternatively, the modified log-likelihood ratio test as described in the copending and commonly assigned U.S. patent application entitled "Weight Based Background Discriminant Functions In Authentication Systems " (U. Chaudhari et al.), filed herewith, may be used instead of the log-likelihood equation (5) set forth above.

Thus, with reference to Figure 2, a speaker verification process 200 may employ
15 various background speaker models 200a, which may be constructed similarly to that indicated at 100 in Figure 1 (*i.e.*, with varying levels of phonetic detail). At the same time, a target speaker's voice print may be obtained at 200b. Model parameters 201a, 201b,

corresponding to background speaker models 200a and 200b, respectively, are then preferably input into the step of pickmax-score calculation and the log-likelihood ratio test at 209. Preferably, also serving as input into test 209 are frames 207 resulting from a test utterance 203, preferably with the intermediary step of feature extraction (205). As
5 discussed above, while many suitable methods exist for undertaking such steps, the processes described in U. Chaudhari et al., *supra*, are believed to be particularly appropriate in this context.

A score 211 (*L*) resulting from test 209 then preferably is input into decision logic (213), with the end result that a (threshold) decision on acceptance or rejection is made at
10 step 215.

It will be appreciated from the foregoing that the technique described hereinabove with reference to Figure 2 creates phonetically structured speaker models. Using the complete unit ensemble provided by the model, a scoring method then assigns the best matching likelihood to each feature vector frame and thus maximizes the resulting model
15 score. This improves the significance of the those models that carry useful information for that particular frame in the verification and thus their "competitiveness" in the final log-likelihood ratio test.

Furthermore, as the score calculation mechanism (verification stage) works on a frame-by-frame basis and picks the maximum likelihood across all phonetic units, there is essentially no need for explicit labeling information during this stage. This may save a considerable amount of computation normally associated with phonetic analysis.

- 5 With reference to Figure 3, the identification of a speaker (*i.e.*, determining identity I [indicated at 320]) based on the test utterance 303 as denoted above and involving the score 311 calculated as (2), or specifically (4), can be carried out as a maximum-likelihood classification:

$$I = \arg \max_{i \leq Y} S(U | M_y)$$

- 10 with Y denoting the total number of speakers enrolled in (*i.e.* known to) the identification system. Other components of the identification system in Figure 3 that are analogous to components in Figure 2 bear reference numerals that are advanced by 100.

- 15 It should be appreciated that the specific task of “identification” can involve recognition methods such as “speaker segmentation” and “speaker tracking”. These tasks will preferably use a likelihood score measure for which the generalized score calculation (2) and its preferred form (4) can be applied. A detailed description of these additional tasks can be found in S. Maes, “Conversational Biometrics,” (Proc. of the

European Conference on Speech Communication and Technology [EUROSPEECH'99],
Budapest, Hungary, 1999).

It should be appreciated that, in contrast to the processes described hereinabove,
conventional techniques typically calculate the speaker scores based on either global
5 (phonetically unstructured) models or on different levels of phonetic detail -- in the latter
case, however, smoothing techniques, e.g. linear interpolation, between models with the
same phonetic distinction (but on different levels of coarseness) are applied, which entails
the necessity of phonetic labeling during test as well as the need for tuning interpolation
constants using additional development data.

10 It is to be understood that the present invention, in accordance with at least one
presently preferred embodiment, includes a target speaker model generator, a receiving
arrangement for receiving an identity claim and a decision arrangement for ascertaining
whether the identity claim corresponds to the target speaker model. Together, the target
speaker model generator, receiving arrangement and decision arrangement may be
15 implemented on at least one general-purpose computer running suitable software
programs. These may also be implemented on at least one Integrated Circuit or part of at
least one Integrated Circuit. Thus, it is to be understood that the invention may be
implemented in hardware, software, or a combination of both.

If not otherwise stated herein, it is to be assumed that all patents, patent applications, patent publications and other publications (including web-based publications) mentioned and cited herein are hereby fully incorporated by reference herein as if set forth in their entirety herein.

- 5 Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.